Model Representation

$x^{(i)}$ – ith input variable or feature (in this case, living area)
$y^{(i)}$ – ith output or target variable that we are trying to predict (price)
$(x^{(i)}, y^{(i)})$ – One training example.

Thus, the dataset we use to learn a function is called a training set because it is basically a list of m training examples $(x^{(i)}, y^{(i)})\ ; i = 1, \dots, m$

Our goal is to use the training set and learn a function h(x) (hypothesis function) that is a good predictor for the corresponding values of y.

If the target variable is continuous, then this learning problem is a regression problem. If discrete, it is a classification problem.

In a simple linear regression, the hypothesis function can be represented by $h_\theta(x) = \theta_0 + \theta_1 x$

Cost Function

This function allows us to simply measure the accuracy of our hypothesis function. It takes a fancier average of the squared difference between the predicted values and actual values.

We can represent the cost function as

$$J(\theta_0\ , \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

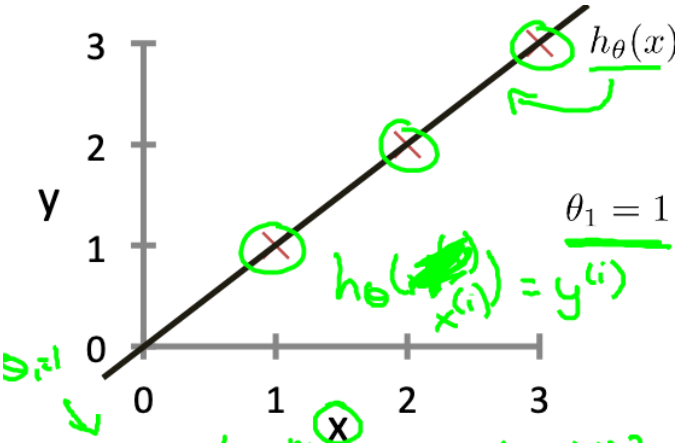
This function is otherwise called the "Squared error function", or "Mean squared error". The goal is to minimize $J(\theta_0\ , \theta_1)$ with respect to $\theta_0\ , \theta_1$.

Cost Function Intuition – I

Let's assume that $\theta_0 = 0$. In this case, $h_\theta(x) = \theta_1 x$, and the cost function would be

$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\theta_1 x^{(i)} - y_i)^2$$

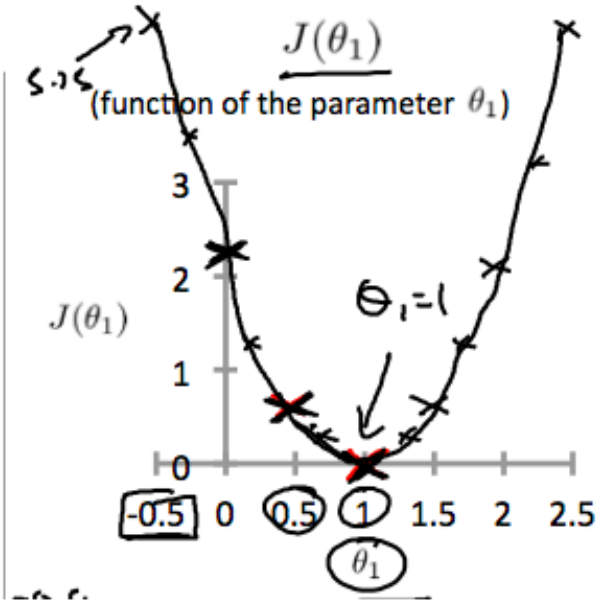
Thus, we want to minimize $J(\theta_1)$ with respect to θ_1.



The graph illustrates that the hypothesis function predicts all the input variables correctly. Thus, the output of the cost function is

$$\frac{1}{2m}[\,(1-1)^2 + (2-2)^2 + (3-3)^2\,]$$
$$\therefore\ J(\theta_1) = 0$$

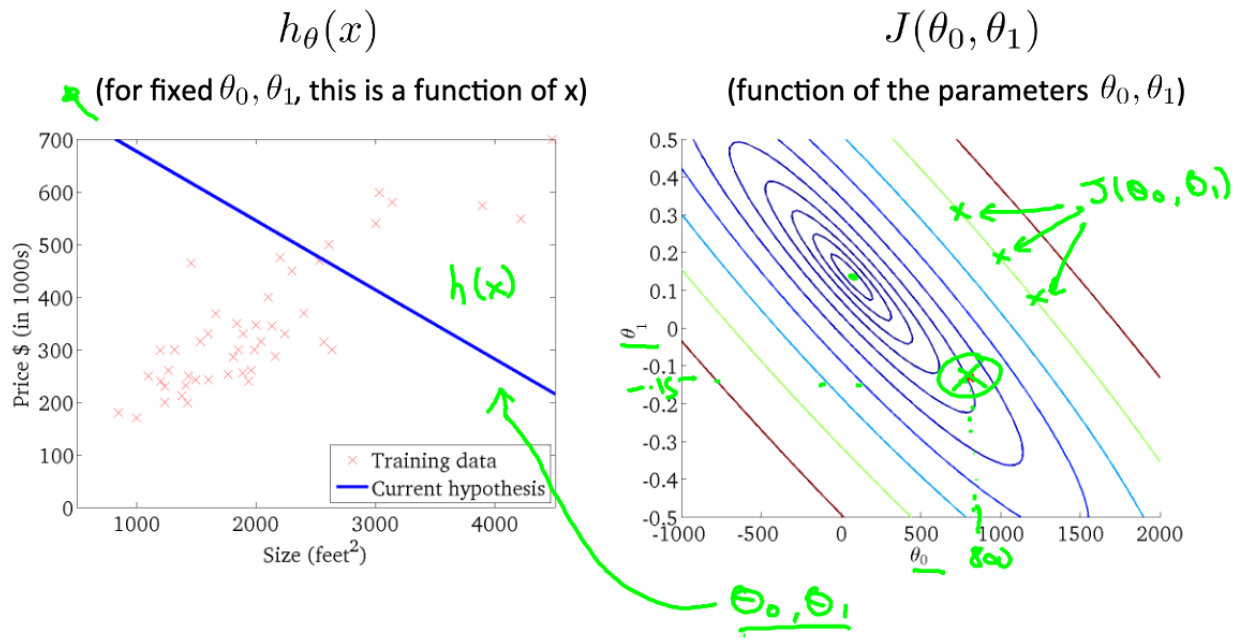
Essentially, we choose a slope for the hypothesis function and calculate the cost function error of that slope.
A different slope yields a different cost as the distances between each prediction and actual value will change. Repeating this process for many different slopes will result in several ordered pairs of slope and that cost function error that we can plot as shown below. This parabola indicates that $\theta_1 = 1$ is the best choice.



Cost Function Intuition – II

In many cases, the y-intercept of the hypothesis function will not be 0. Thus, we will have to minimize $J(\theta_0, \theta_1)$ with respect to θ_0 *and* θ_1. However, the process from Intuition I still applies: Use the training set to learn a hypothesis function, calculate the cost function error using the cost function, repeat the process using several different $\theta_0's$ *and* $\theta_1's$, and choose the pair that minimizes the cost function.

We can use a contour figure to visualize the cost function errors for different $\theta_0's$ *and* $\theta_1's$.



The three green points found on the green line above have the same value for $J(\theta_0, \theta_1)$. Reducing this error gets us closer to the center of the contour plot.